TRANSPOSON INSERTION FOLLOWED BY SEQUENCING METHOD TO STUDY INTERACTIONS BETWEEN GENES

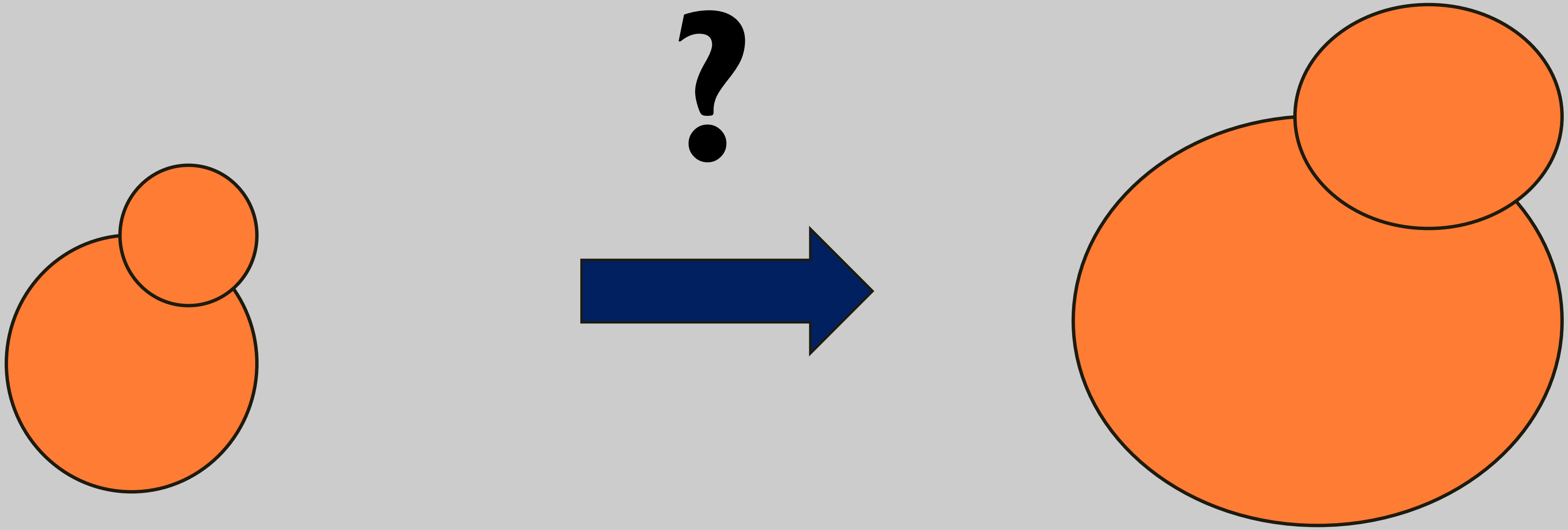
Alena Martsul

May 24, 2018

Why is it important to study interactions between genes?

Usually multiple genes contribute to one specific phenotype (trait)

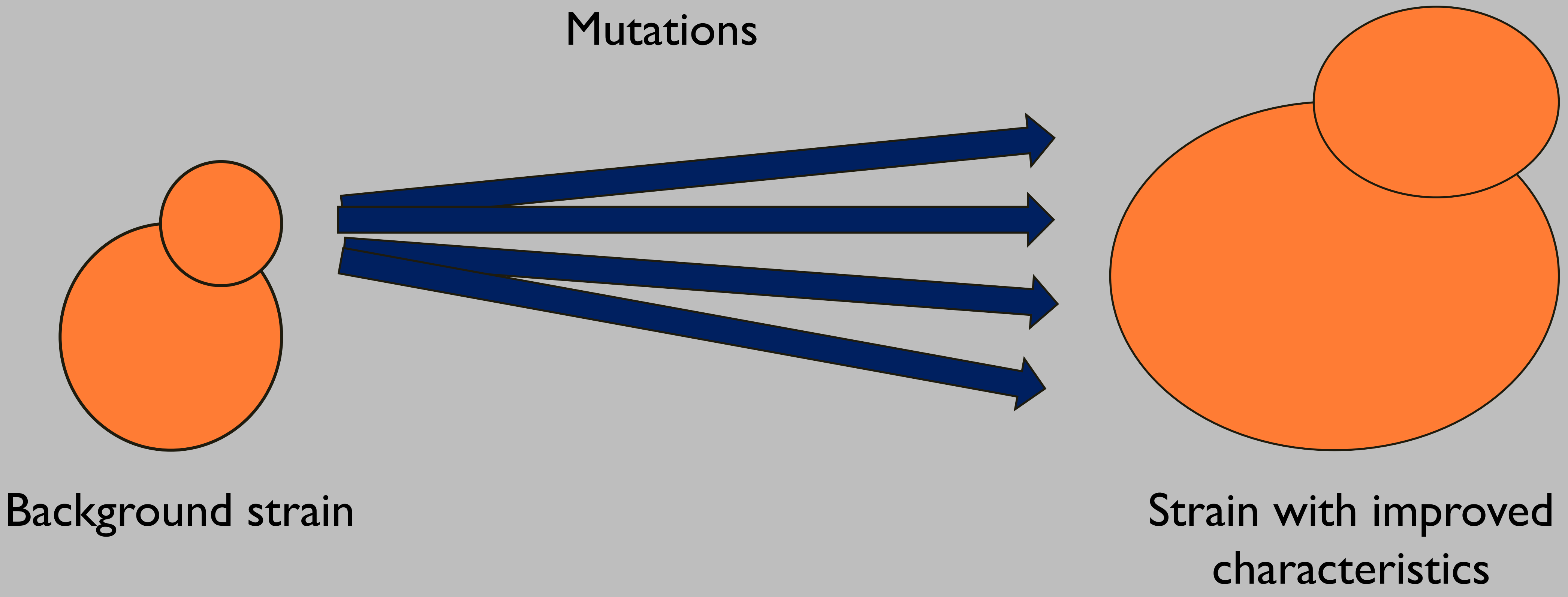
Example: You want to construct a yeast strain that grows fast in medium with low pH



Background strain

Strain with improved characteristics

Example: You want to construct a yeast train that grows fast in medium with low pH



How can we study interactions between mutations?

Choosing experimental approach

Requirements:

- Simple and reproducible method to introduce mutations
- A method to distinguish mutations and track their frequencies over time
- Reasonable timeline

Choosing experimental approach

Requirements:

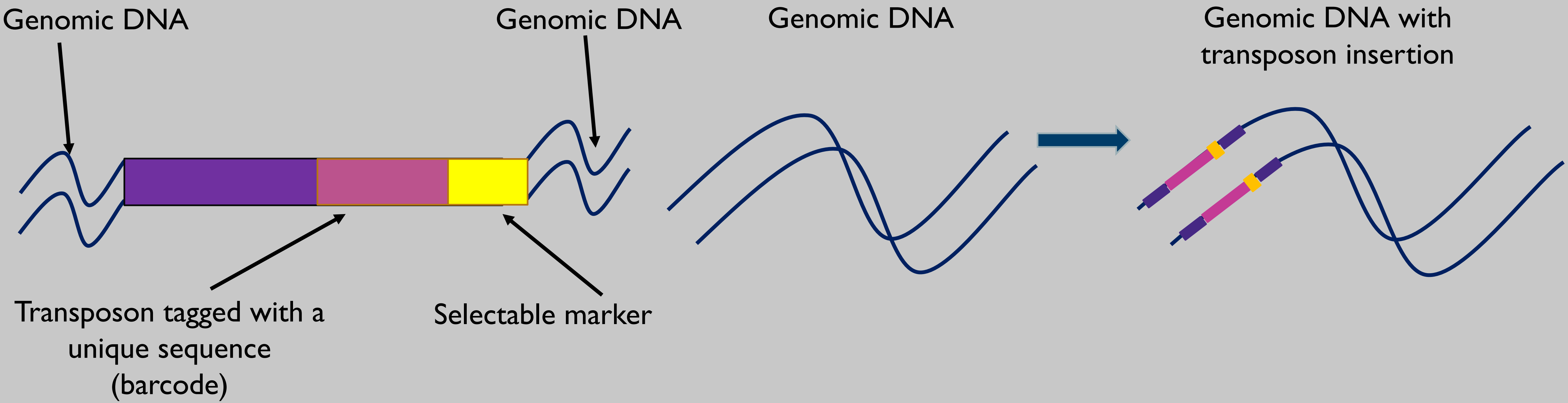
- Simple and reproducible method to introduce mutations
- A method to distinguish mutations and track their frequencies over time
- Reasonable timeline

Alternative approaches:

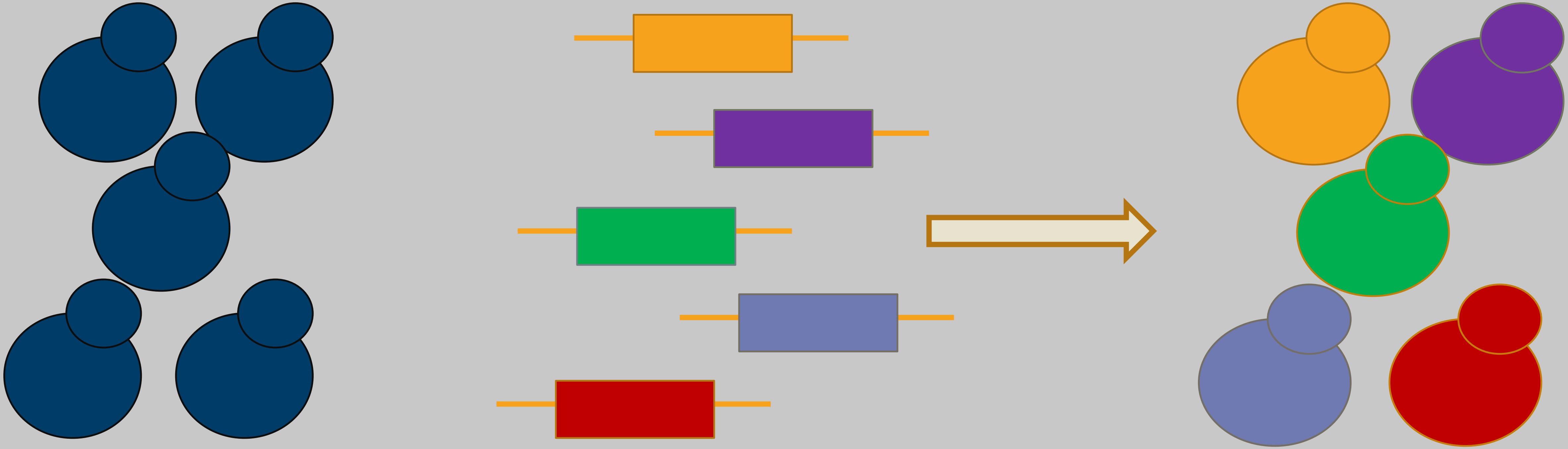
- Sampling from the wild
- Mutagenesis (chemical, UV exposure, etc.)
- Use of yeast deletion collections

Transposon mutagenesis followed by sequencing (TnSeq) method

1. Transposon insertion


Genomic DNA
Genomic DNA
Genomic DNA
Genomic DNA with transposon insertion
Transposon tagged with a unique sequence (barcode)
Selectable marker

2. Generation of mutant library

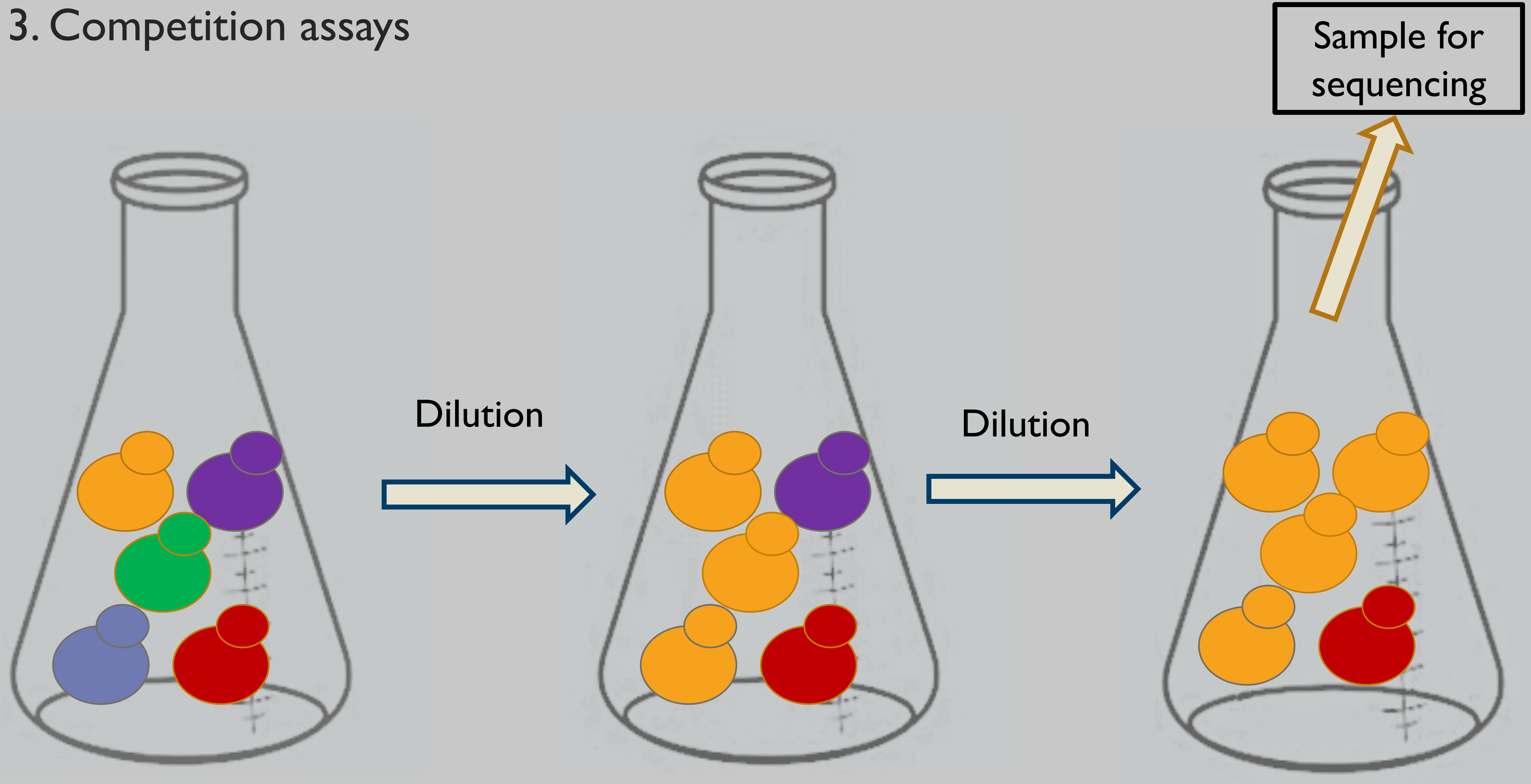


Yeast or bacterial strain

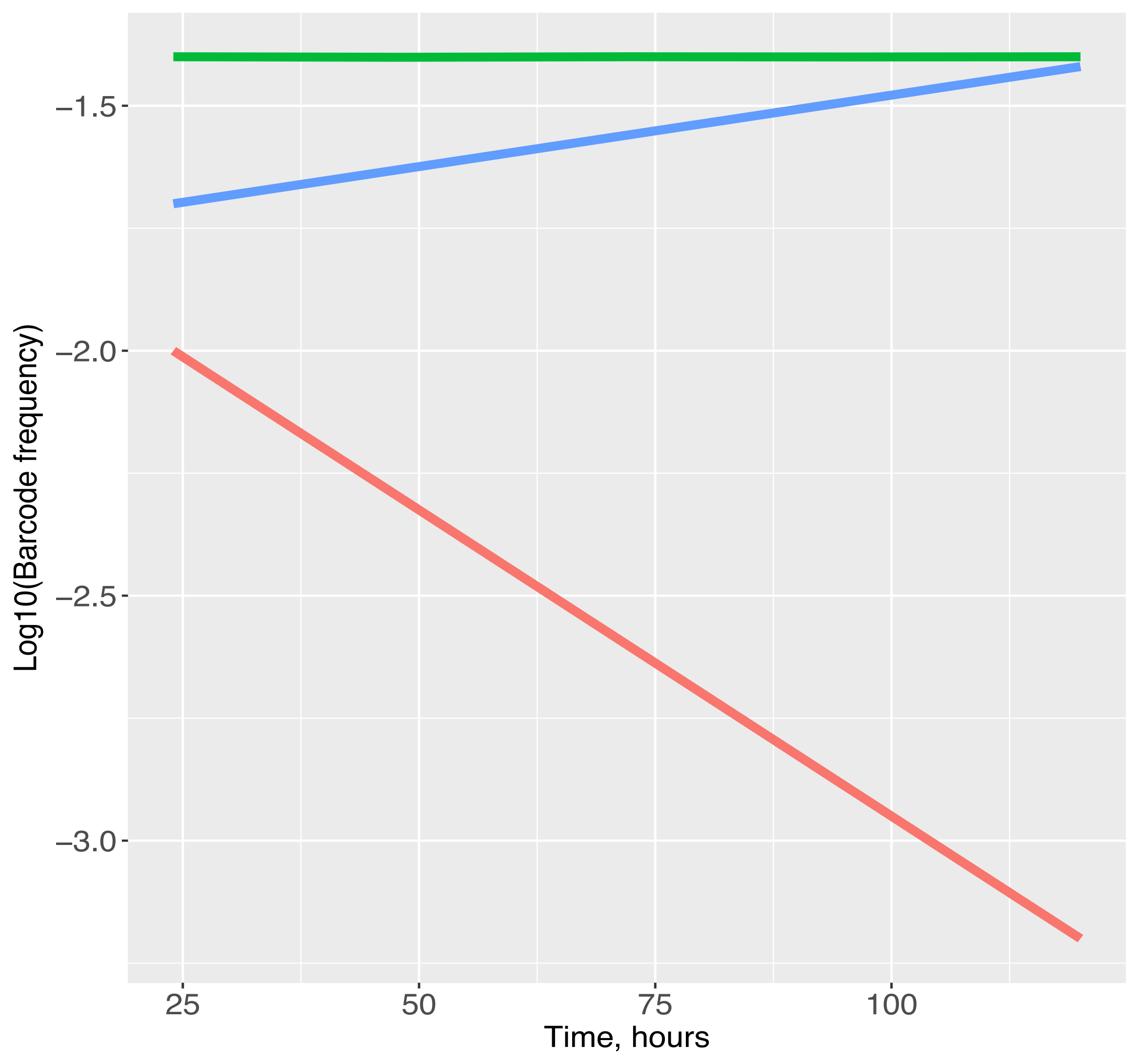
Transposon library

Mutant library

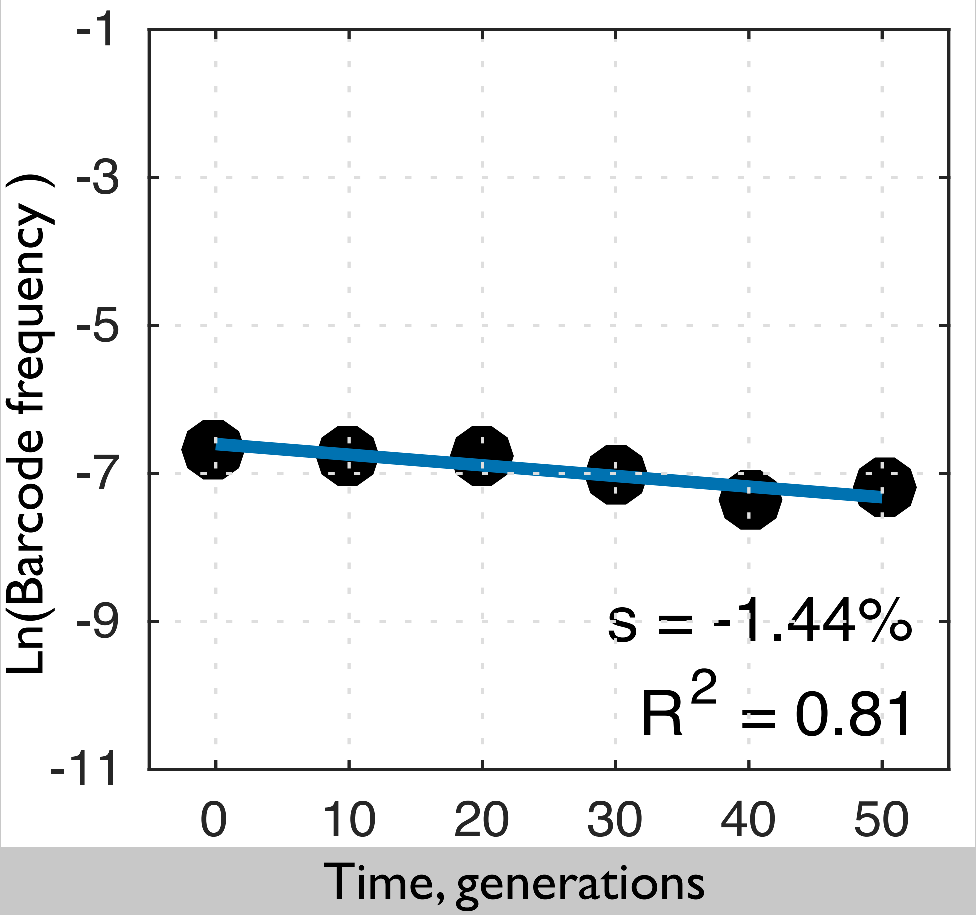
3. Competition assays



4. Tracking barcode frequency trajectories



5. Estimating fitness effects of mutations

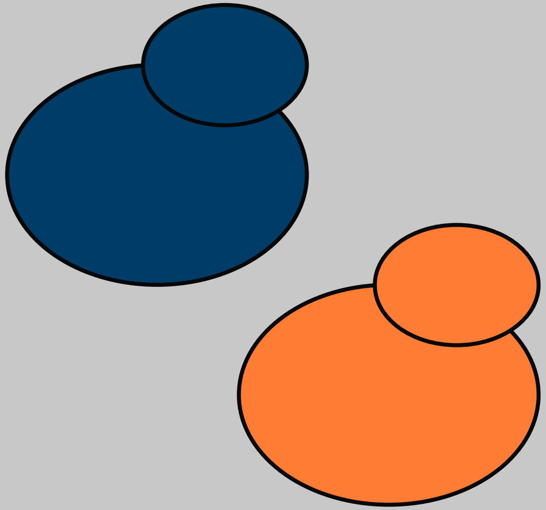


LET’S PRACTICE!

Dataset description

Model organism:

2 *Saccharomyces cerevisiae* strains that differ by 5 mutations



Dataset description

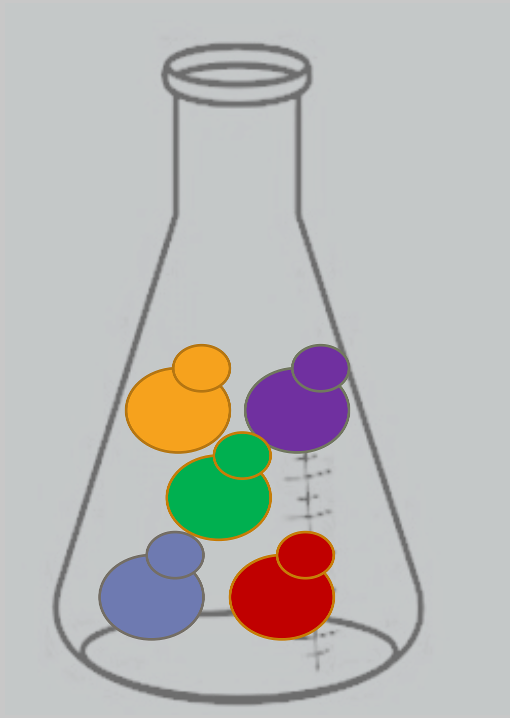
Mutations tested:

~1000 mutations with different genomic location

Dataset description

Environment:

Synthetic complete medium with low pH (3.0)

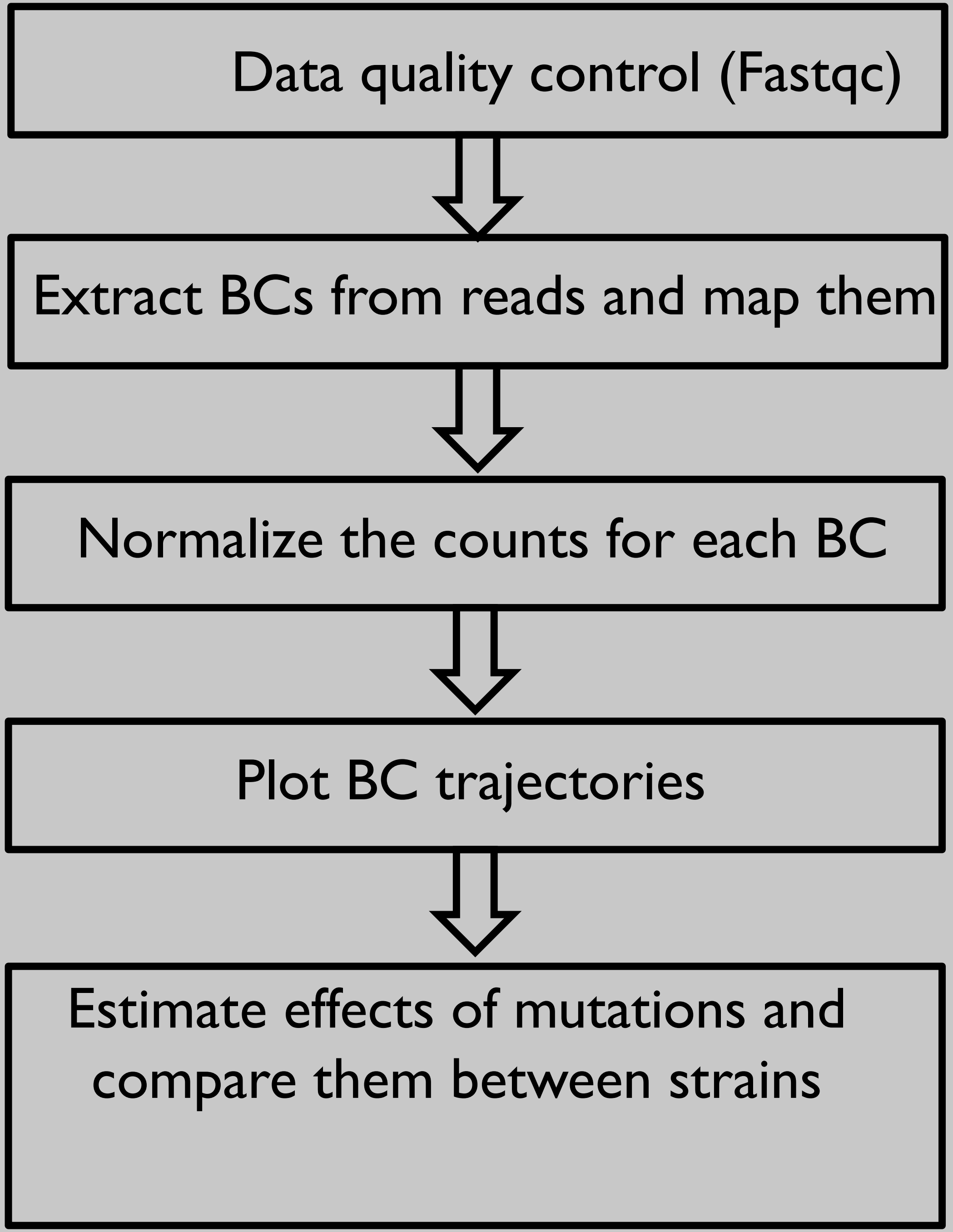


Questions we want to answer

Do some of the mutations have different fitness effect?

If yes, how different this effect can be?

Data analysis workflow



Data quality control (Fastqc)

Extract BCs from reads and map them

Normalize the counts for each BC

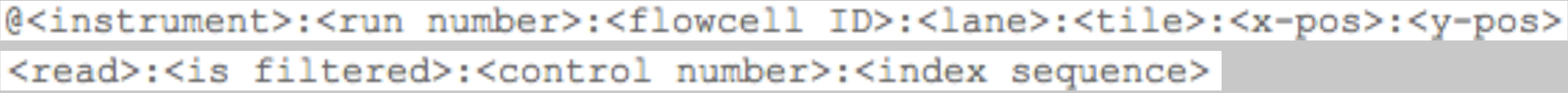
Plot BC trajectories

Estimate effects of mutations and compare them between strains

What information do fastq files contain?

```
@NS500672:54:HL775BGXX:1:11101:22716:1042 1:N:0:CCCCGG
CCGCCNATGCCCATGCCACAGTTGTTGAGCTTGAGTTCCTGCAGGGTGAAGCAGGCTGAGCTCTTGA
GCAGGGCCTCGAA
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEE
```

First line is the information about the location of the read and specific sequencing machine used:



```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>
<read>:<is filtered>:<control number>:<index sequence>
```

Second line is the nucleotide sequence called

Third line is "+" and can optionally be followed by a repeat of the filename in line 1

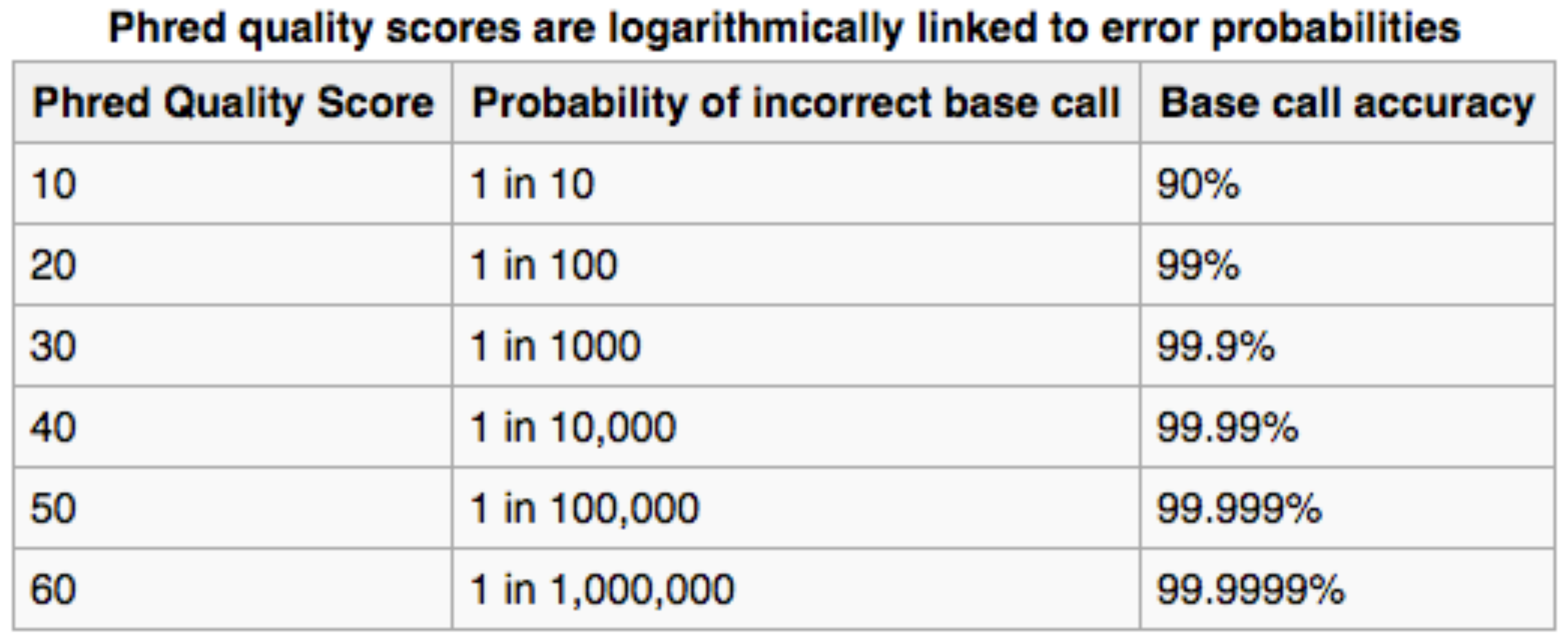
Fourth line contains the quality score as determined by the sequencer

How can we check the quality of sequencing data?

Fastq File – Phred Quality Score

$$\boxed{Q= -10Log_{10}P}$$

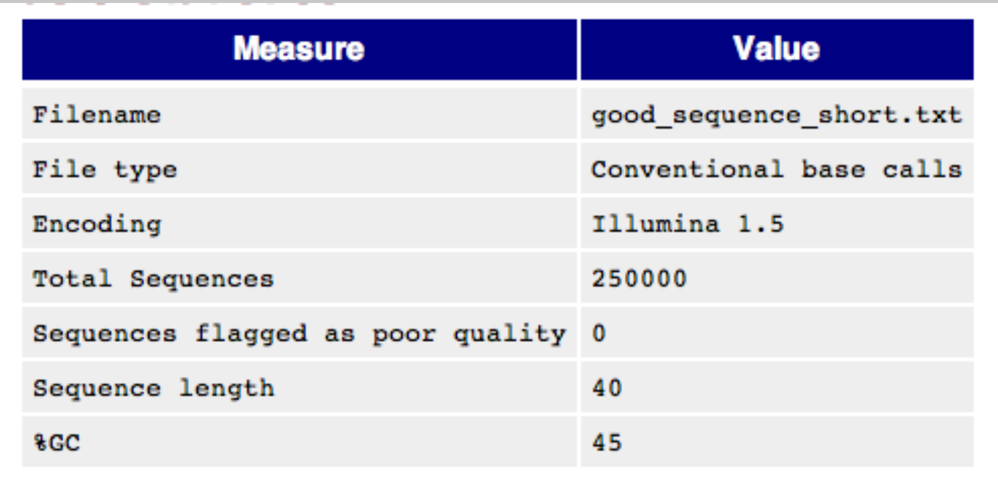
Quality scores report the probability that the base call is incorrect



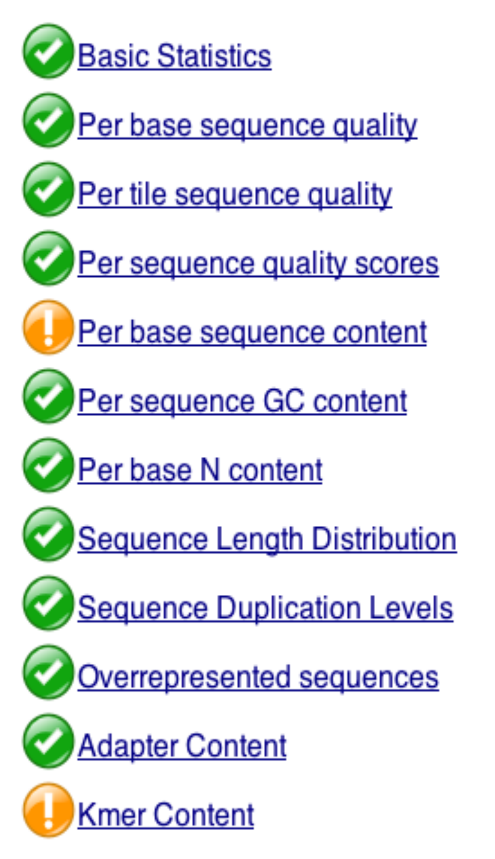
Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

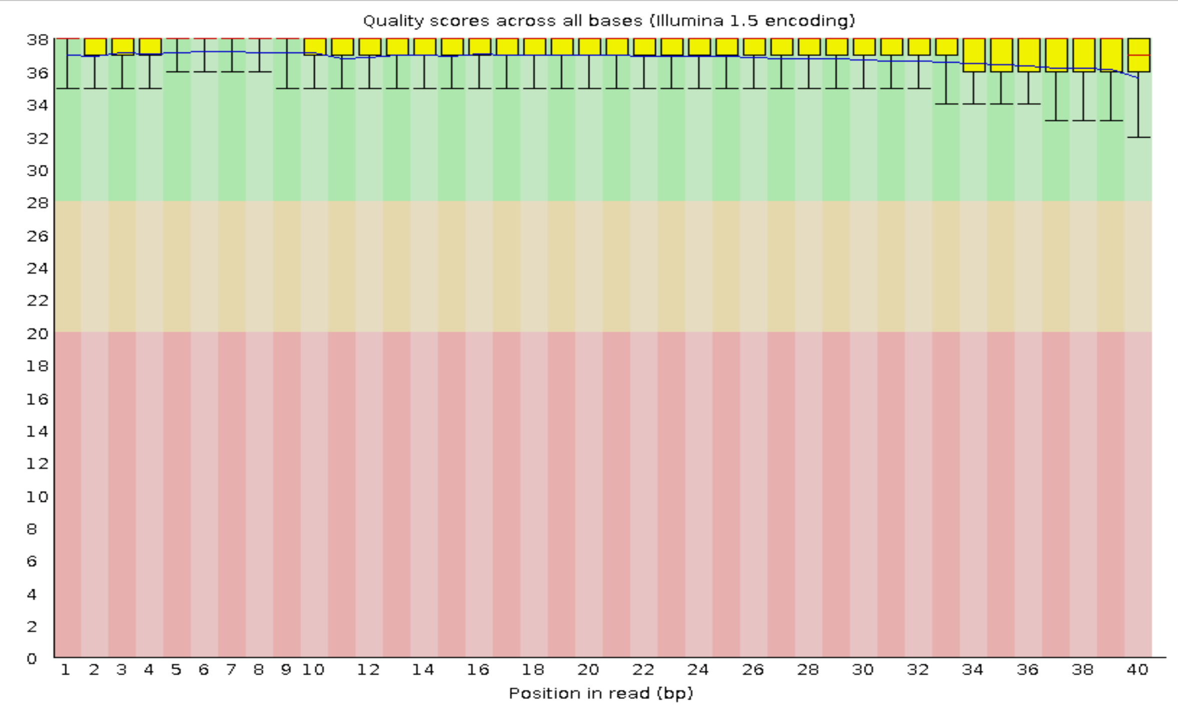
Field standard is to accept bases with quality >20

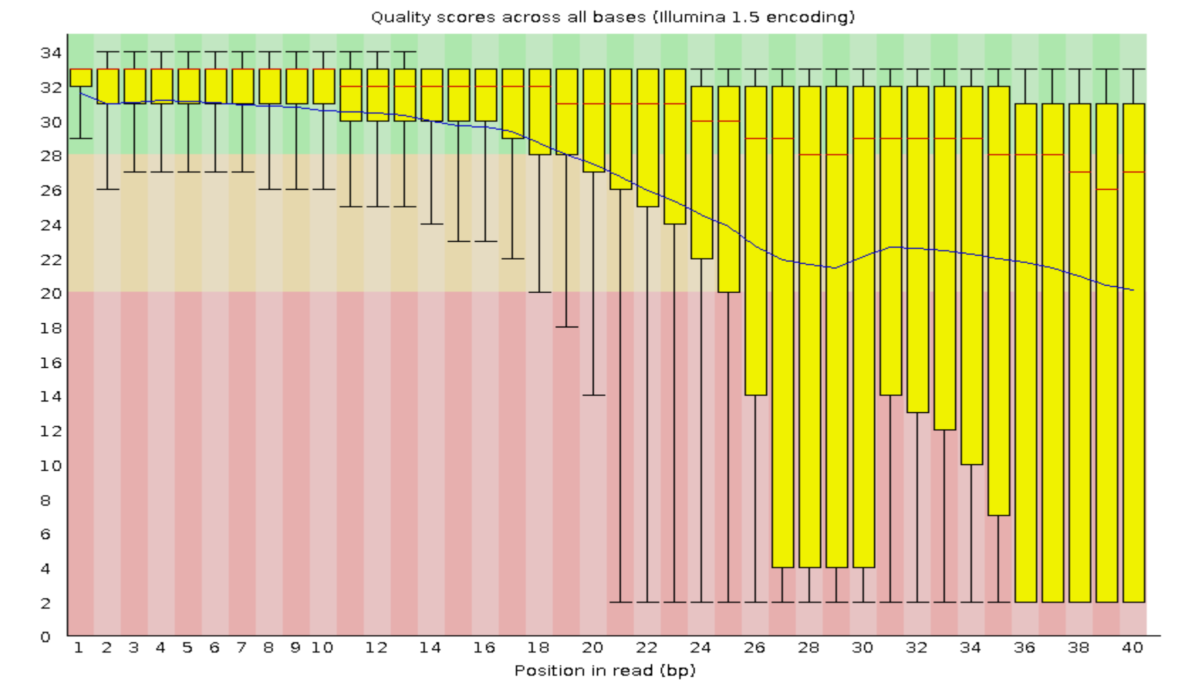


Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

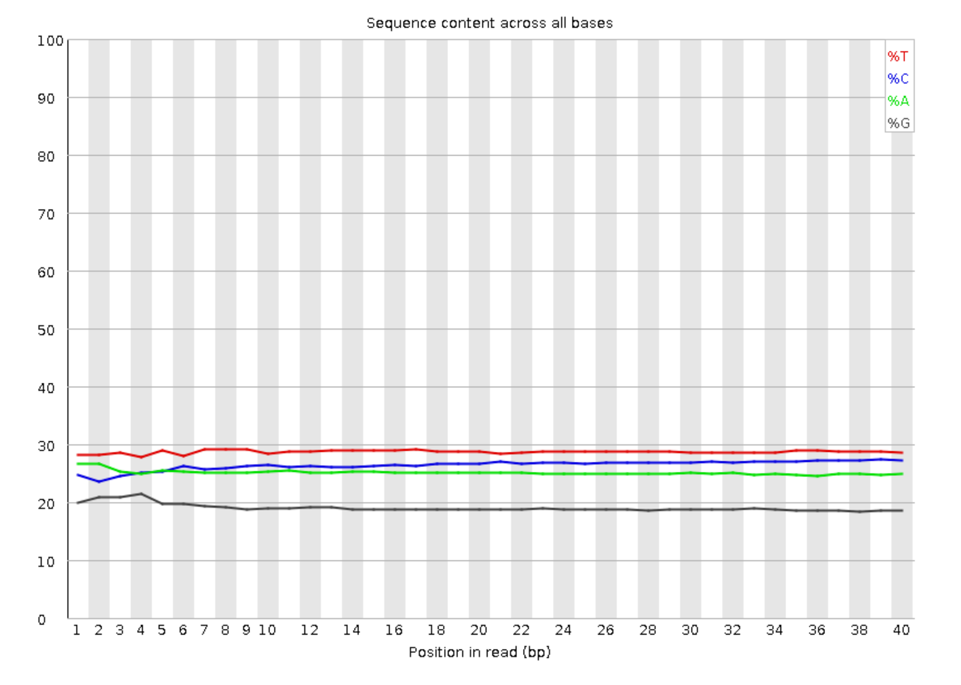


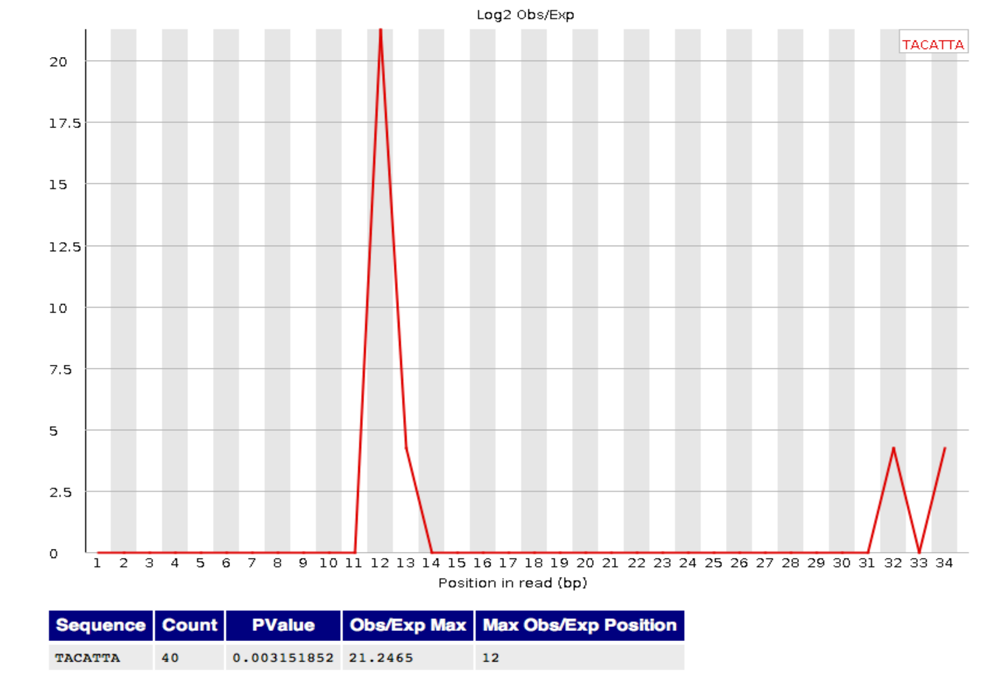
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content



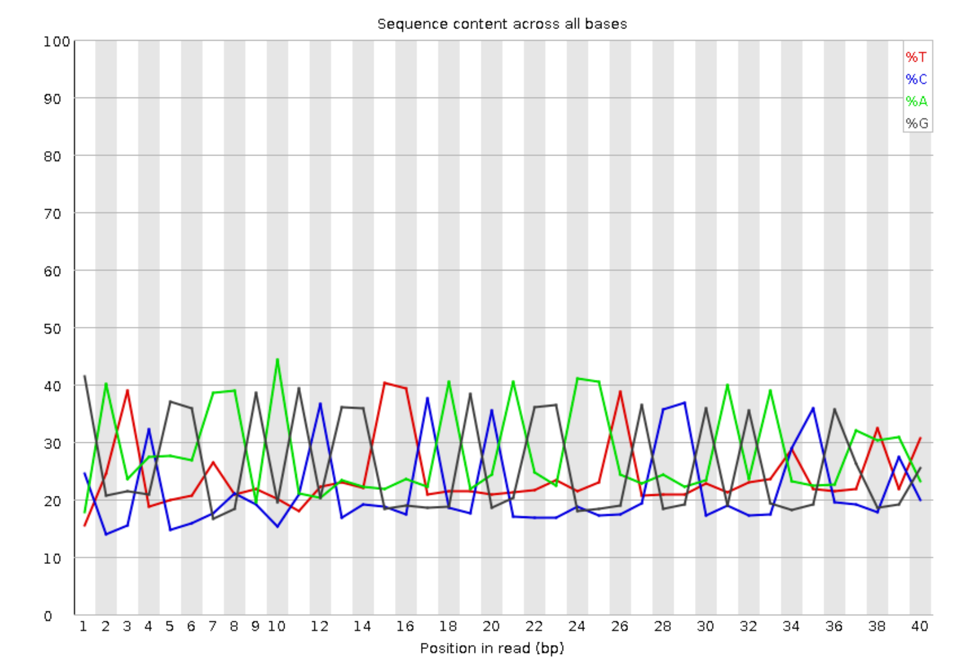


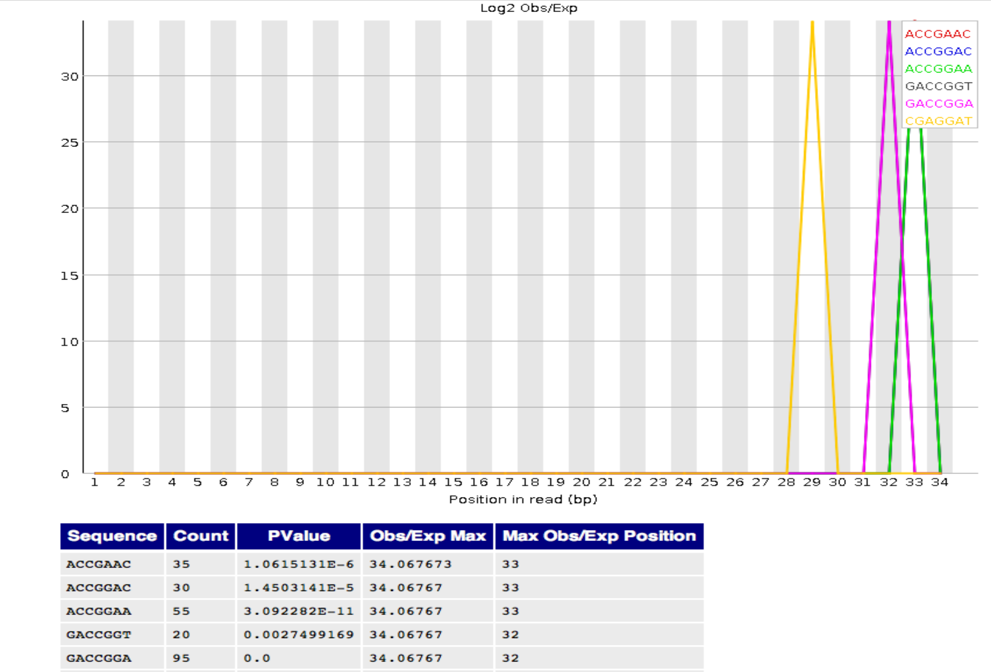
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/





Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TACATTA	40	0.003151852	21.2465	12

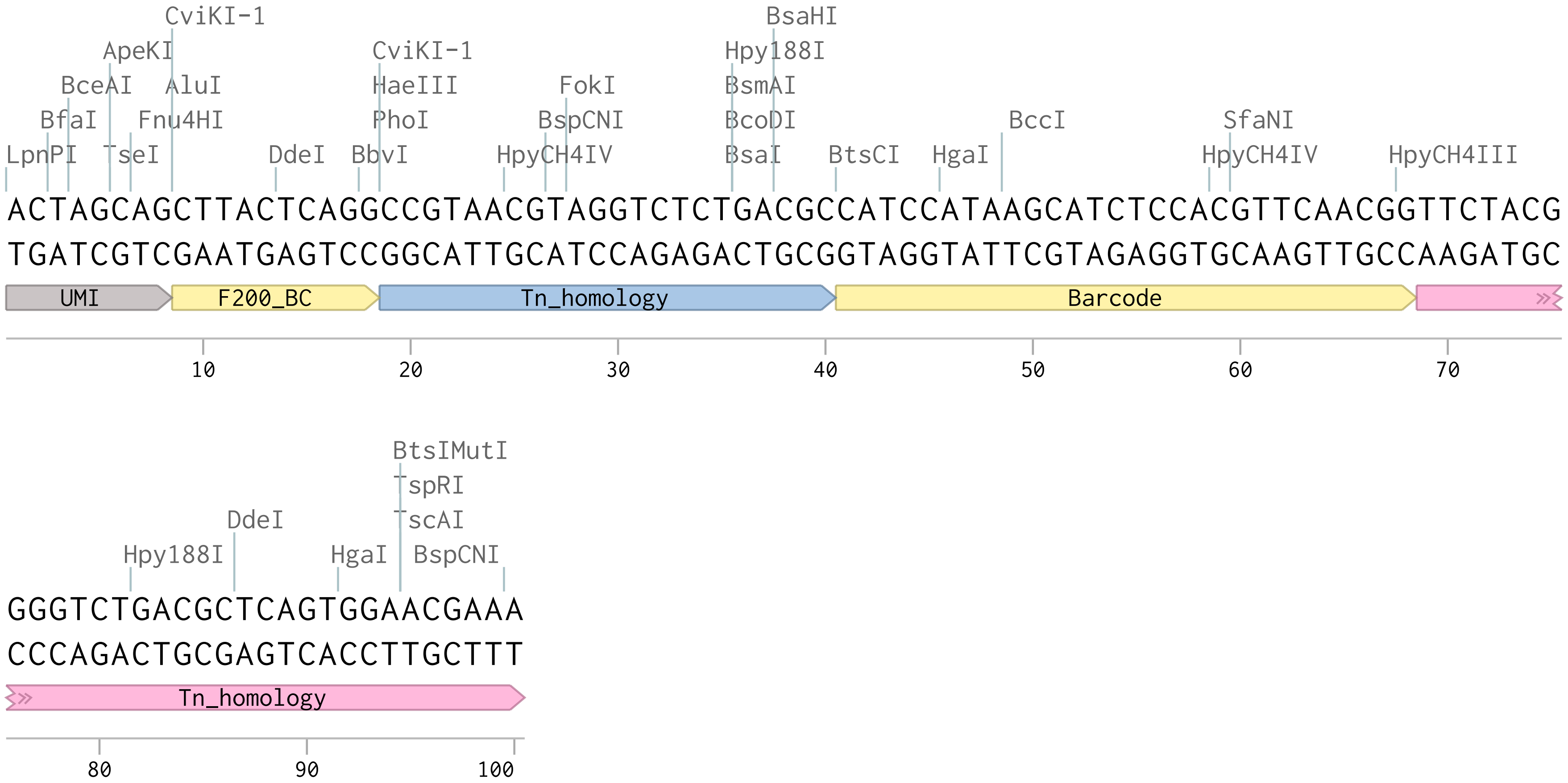




Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
ACCGAAC	35	1.0615131E-6	34.067673	33
ACCGGAC	30	1.4503141E-5	34.06767	33
ACCGGAA	55	3.092282E-11	34.06767	33
GACCGGT	20	0.0027499169	34.06767	32
GACCGGA	95	0.0	34.06767	32

Barcode extraction procedure and mapping

DivAnc_F200 (100 bp)



Can we use barcode raw counts for the analysis?

LETS CODE!

Please fill out the assessment below:

https://docs.google.com/forms/d/e/1FAIpQLScenZBfkADH6dgbvTYfoNi5LbvGB4I7AgdIhGr3ey_IhSKQYQ/viewform?usp=sf_link

Thank you!